Study Protocol

Study Title

The Effect of Prompt Framing on AI Recommendations for Low-Dose Aspirin in Pregnancy

1. Research Question

Does changing the framing and wording of the same medical question affect the accuracy and completeness of recommendations generated by ChatGPT and Gemini regarding which pregnant patients should receive low-dose aspirin for preeclampsia prevention?

2. Study Objective

The primary objective is to determine whether prompt framing changes how accurately and completely AI systems communicate current recommendations about eligibility for low-dose aspirin during pregnancy.

The secondary objectives are to evaluate:

- Whether the AI systems provide consistent answers when the same prompt is repeated.
- Whether they accurately represent differences among ACOG and SMFM including areas of agreement and legitimate disagreement
- Whether any citations provided are real, relevant, and supportive of the claims made.
- Whether they make claims that cannot be supported by the evidence identified during the study.

3. Study Design

This will be a prospective, structured comparative evaluation of AI-generated medical recommendations.

Two AI systems will be tested:

- ChatGPT
- Gemini

Each system will receive three prompts. Each prompt will be tested three times.

The study will therefore include:

- Two AI systems
- Three prompt conditions
- Three repetitions per AI–prompt condition
- Eighteen total AI responses

4. Study Variables

Independent Variables

1. **AI system**
    - ChatGPT
    - Gemini
2. **Prompt framing**
    - Neutral
    - Formal clinical
    - Evidence-focused

Primary Outcomes

1. Accuracy of medically relevant claims
2. Completeness of the recommendation

Secondary Outcomes

1. Appropriate representation of guideline agreement or disagreement
2. Unsupported or unverifiable claims
3. Citation validity
4. Consistency across repeated responses
5. Response time

Response time will be exploratory and will not be treated as a measure of medical quality.

Phase 1: Development of the Evidence Reference

5. Purpose of the Reference Evidence

Before AI testing begins, an evidence reference will be developed using authoritative medical recommendations and the research supporting them.

The evidence reference is not intended to create one artificially absolute answer when legitimate medical disagreement exists. Instead, it will identify:

- Areas where major sources agree
- Areas where recommendations differ
- Claims that are strongly supported
- Claims that remain uncertain
- Claims for which insufficient evidence is available

The AI responses will then be judged according to whether they represent this evidence accurately and fairly.

6. Evidence Sources

The investigator will review:

- Current clinical guidelines from major professional organizations
- Governmental or national recommendations
- Systematic reviews and meta-analyses
- Evidence reviews used to develop guidelines
- Original randomized controlled trials cited by reviews or guidelines

The primary professional-organization sources will be the American College of Obstetricians and Gynecologists (ACOG) and the Society for Maternal-Fetal Medicine (SMFM). Additional sources may be used to trace their recommendations to systematic reviews, meta-analyses, evidence reviews, and original randomized controlled trials.

General health websites, news articles, blogs, social media posts, and AI-generated summaries may be examined to understand how information spreads, but they will not determine the evidence reference unless their claims can be traced to stronger evidence.

Because ACOG and SMFM frequently publish aligned recommendations, agreement between them will not be interpreted as independent confirmation but rather as consistency between the professional organizations reviewed.

7. Reference Questions

The following questions will be answered before AI responses are scored:

Eligibility

1. Which patients are considered at high risk for preeclampsia?
2. Is one high-risk factor sufficient for an aspirin recommendation?
3. Which factors are considered moderate risk?
4. How many moderate-risk factors are required before aspirin is recommended or considered?
5. Do organizations define high- and moderate-risk factors differently?
6. Do organizations differ in how they address:
    - First pregnancy
    - Obesity
    - Maternal age
    - Family history
    - Previous adverse pregnancy outcomes
    - Multifetal pregnancy
    - Chronic hypertension
    - Diabetes
    - Kidney disease
    - Autoimmune disease
    - Socioeconomic or demographic factors

Treatment Details

7. What aspirin dose or dose range is recommended?
8. At what gestational age should aspirin be started?
9. Is there a preferred initiation period?
10. What guidance exists for patients who present after the preferred period?
11. When should aspirin be discontinued?
12. What contraindications, cautions, or qualifications are relevant?

Evidence and Disagreement

13. Which recommendations are supported by randomized controlled trials?
14. Which recommendations are supported by systematic reviews or meta-analyses?
15. Which recommendations rely mainly on expert interpretation or indirect evidence?
16. Where do major medical organizations agree?
17. Where do they meaningfully differ?
18. Are differences caused by:

- Different evidence
- Different definitions
- Different risk thresholds
- Different healthcare systems
- Different judgments about benefits and harms

19. Has newer evidence changed or challenged older recommendations?
20. What limitations exist in the evidence?

Phase 2: AI Testing

9. AI Systems

The following systems will be tested:

- ChatGPT
- Gemini

For each trial, the data collector will record:

- AI system
- Model name or version displayed
- Account type, if relevant
- Date of testing
- Local submission time
- Local completion time
- ~~Prompt condition~~
- ~~Repetition number~~
- Response time in seconds

Why Date and Time Will Be Recorded

AI systems can change because of model updates, system settings, or access to newer information. Recording the date, time, and model identifies the conditions under which each answer was produced.

Testing will be completed within the shortest practical period to reduce the likelihood of model changes occurring during data collection.

10. Standardized Prompts

Prompt 1: Neutral

> Which pregnant patients should receive low-dose aspirin to prevent preeclampsia, and what are the recommended dose and timing?

Prompt 2: Formal Clinical

> Which pregnant patients are appropriate candidates for low-dose aspirin for the prevention of preeclampsia, and what dose and initiation timing are recommended?

Prompt 3: Evidence-Focused

> Based on current evidence-based clinical recommendations, which pregnant patients should receive low-dose aspirin for preeclampsia prevention, and what dose and initiation timing are recommended?

These prompts ask the same central medical question but differ in framing and terminology.

Because the third prompt explicitly refers to evidence-based recommendations, this study evaluates **prompt framing**, not only minor word substitution.

Phase 3: Randomized Data Collection

11. Randomized Testing Schedule

The 18 trials will be listed in a spreadsheet.

Each trial will include:

- AI system
- Prompt condition
- Repetition number

A random number will be assigned to each trial using a spreadsheet function. The trials will then be sorted according to the random numbers.

The finalized randomized schedule will be saved before data collection and followed exactly.

Randomization is intended to prevent one AI system or prompt condition from being tested only at one particular time or in one continuous block.

12. Data Collector

A second person will conduct the AI testing.

The investigator who will score the responses will not observe the testing process.

The data collector will receive:

- The three finalized prompts
- The randomized schedule
- Trial-recording instructions
- Technical-failure rules
- A private identification-key template

13. Procedure for Each Trial

For every trial, the data collector will:

1. Confirm the assigned AI system.
2. Confirm the model, when displayed.
3. Start a new conversation with no previous chat context.
4. Copy and paste the assigned prompt exactly.
5. Submit the prompt once.
6. Record the submission time.
7. Allow the response to finish without interruption.
8. Record the completion time.
9. Save only the first response.
10. Not regenerate the answer.
11. Not ask follow-up questions.
12. Not correct or guide the AI.
13. Copy the complete response without changing its wording.
14. Save a screenshot.
15. Record any citations or links.
16. Record the trial details in the private identification key.

If the AI asks a clarifying question rather than answering, that will be retained as the first response.

A response will not be repeated because it appears incorrect, incomplete, short, unusual, or disappointing.

14. Technical Failures

A trial may be replaced only when:

- The internet connection fails before the response is completed.
- The AI platform produces a technical error.
- The wrong prompt is accidentally submitted.
- The wrong AI system or model is used.
- The full response cannot be saved.

The failed trial and reason for replacement will be documented.

Phase 4: Blinding

15. Preparation of Responses

After all trials are complete, the data collector will:

1. Remove AI platform names.
2. Remove model labels and logos.
3. Remove obvious platform interface formatting.
4. Preserve the exact response wording.
5. Assign a random code to each response, such as R01–R18.
6. Place all responses into one randomized document.
7. Keep the identification key private.

The investigator will not know:

- Which AI generated each response
- Which prompt produced it
- Which repetition it represents
- Which responses belong to the same condition

The identification key will be revealed only after all scoring is completed and recorded.

Complete blinding may not be possible if writing style or citation formatting suggests the AI source. This will be documented rather than ignored.

Phase 5: Outcome Assessment

16. Claim Identification

Each response will first be divided into medically relevant claims.

A medically relevant claim is a statement concerning:

- Eligibility for aspirin
- Risk factors
- Number or combination of risk factors
- Timing
- Dose
- Discontinuation
- Contraindications
- Guideline recommendations
- Benefits or harms
- Strength or certainty of evidence

General statements that do not affect the medical recommendation will not be scored for accuracy.

17. Claim Classification

Each medically relevant claim will be classified as one of the following:

Correct

The claim is consistent with the evidence reference.

Partially Correct

The main idea is supported, but the claim is incomplete, overly broad, imprecise, or missing an important qualification.

Incorrect

The claim conflicts with the evidence reference or inaccurately represents an authoritative source.

Uncertain or Disputed

Authoritative sources disagree, or the available evidence does not support one definitive conclusion.

The AI will not be penalized for accurately describing a legitimate disagreement.

Not Addressed by the Initial Reference Evidence

The claim was not covered by the initial reference table.

This category will trigger the additional-verification procedure described below. It will not automatically be classified as incorrect.

18. Additional-Verification Procedure

If an AI response makes a medically relevant claim that is not addressed by the initial reference evidence:

1. The claim will be recorded exactly.
2. The investigator will search the same predefined evidence categories used in Phase 1.
3. Priority will be given to:
    - Current guidelines
    - Systematic reviews
    - Evidence reviews
    - Original studies
4. The new source and search pathway will be documented.
5. The claim will then be classified as:
    - Correct
    - Partially correct
    - Incorrect
    - Uncertain
    - Still unverifiable

Newly identified evidence will not be used selectively for only one AI. If it changes the reference evidence, the updated rule will be applied to every response containing the same type of claim.

This prevents an AI from being penalized merely because it introduced information not anticipated during Phase 1.

19. Primary Outcome: Claim-Level Accuracy

The primary accuracy outcome will be the proportion of scorable medical claims that are correct.

Claims classified as uncertain, disputed, or still unverifiable will be reported separately and excluded from the main accuracy denominator.

The calculation will be:

Accuracy proportion = Number of correct claims ÷ Number of scorable claims

A secondary count will report partially correct claims separately.

Each response will therefore have:

- Number of correct claims
- Number of partially correct claims
- Number of incorrect claims
- Number of uncertain or disputed claims
- Number of unverifiable claims
- Accuracy proportion

20. Primary Outcome: Completeness

Completeness will measure whether the response contains the essential information needed to answer the prompt.

The following five elements will be assessed:

1. High-risk eligibility criteria
2. Moderate-risk eligibility criteria
3. Number or combination of moderate-risk factors
4. Timing of initiation
5. Dose or dose range

Each element will receive:

- **1:** Adequately included
- **0.5:** Mentioned but materially incomplete
- **0:** Missing or incorrect

The completeness score will range from 0 to 5.

A percentage may also be calculated:

Completeness percentage = Points earned ÷ 5 × 100

21. Guideline Representation

This outcome will assess whether the response accurately communicates the existence and nature of guideline recommendations.

Classification:

- **Accurate:** Guideline recommendations are represented correctly.
- **Partially accurate:** Guidelines are mentioned but oversimplified or incompletely described.
- **Inaccurate:** A guideline is misrepresented.
- **Not mentioned:** No guideline is identified.

Because the prompts do not all explicitly require named guidelines, failure to name an organization will not automatically reduce the main accuracy score.

22. Treatment of Uncertainty

Responses will be evaluated for whether they appropriately qualify the recommendation.

Classification:

- **Appropriate qualification:** The answer explains that eligibility depends on risk factors, clinical context, or guideline variation.
- **Limited qualification:** Some caution is present but important uncertainty is omitted.
- **Overgeneralized:** The answer presents aspirin as universally appropriate or inappropriate without necessary context.

23. Citation Evaluation

Citation provision will be analyzed separately because the prompts do not explicitly require citations.

For every cited source, the investigator will determine:

1. Does the source exist?
2. Is it identifiable?
3. Is it authoritative or peer reviewed?
4. Does it directly support the associated claim?
5. Is it represented accurately?
6. Is it:
    - A guideline
    - A systematic review or meta-analysis
    - An original study
    - An evidence review

    - A professional organization webpage
    - Another type of source
7. Does it cite original evidence, or does it only repeat another secondary source?

Citation categories:

- Verifiable and supportive
- Verifiable but only indirectly supportive
- Verifiable but not supportive
- Misrepresented
- Unverifiable

24. Consistency

After blinded individual scoring is complete and the identification key is revealed, the three repetitions within each AI–prompt condition will be compared.

Consistency will be assessed using:

- Variation in accuracy proportions
- Variation in completeness scores
- Changes in eligibility recommendations
- Changes in timing or dose recommendations
- Major contradictions among repetitions

Consistency will be described rather than reduced to one arbitrary score.

25. Response Time

Response time will be measured in seconds from prompt submission until the full answer is completed.

Response time will be reported as an exploratory technical measure only.

Phase 6: Analysis

26. Statistical Analysis Plan

Because there are only three repetitions within each AI–prompt condition, the study will use descriptive analysis.

No statistical significance testing will be performed.

Risk ratios, odds ratios, and inferential confidence intervals will not be used because the small sample would produce unstable and potentially misleading estimates.

27. Response-Level Results

For every response, report:

- Accuracy proportion
- Number of partially correct claims
- Number of incorrect claims
- Number of uncertain or unverifiable claims
- Completeness score
- Guideline-representation category
- Uncertainty category
- Number and type of citations
- Response time

28. AI-Level Comparisons

For ChatGPT and Gemini, report:

- Mean and median accuracy proportion
- Range of accuracy proportions
- Mean and median completeness score
- Number of incorrect claims
- Number of unverifiable claims
- Number of responses containing overgeneralized recommendations
- Number of citations provided
- Percentage of citations that were verifiable and supportive

29. Prompt-Level Comparisons

For each prompt condition, report:

- Individual response results
- Mean and median accuracy proportion
- Mean completeness score

- Range of results
- Frequency of incorrect or overgeneralized recommendations
- Frequency of valid citations

30. Presentation of Results

The final poster may include:

- A study-flow diagram
- A table comparing the three prompts
- A table comparing ChatGPT and Gemini
- A graph of mean completeness scores
- A graph of claim-level accuracy
- A citation-verification chart
- Examples of important correct, incorrect, or unverifiable claims
- An evidence trail showing:
    - AI claim
    - Cited source
    - Guideline or review
    - Original research

Phase 7: Quality-Control Measures

31. Measures to Reduce Bias and Error

The study will use the following safeguards:

- The protocol will be finalized before AI testing.
- Prompts will not be changed after data collection begins.
- The evidence reference will be created before AI responses are scored.
- The same prompts will be used for both AI systems.
- The trial order will be randomized.
- A separate person will collect the AI responses.
- Only first responses will be retained.
- All completed responses will be included.
- The scorer will be blinded to AI and prompt identity.
- Medically relevant claims will be evaluated using predefined categories.
- Claims not covered by the initial evidence reference will undergo a documented additional-verification process.

- Legitimate guideline disagreement will not automatically be scored as an error.
- Citation validity will be evaluated separately from medical accuracy.
- Scores will be finalized before the identification key is revealed.

32. Protocol Deviations

Any departure from the protocol will be documented, including:

- Model changes
- Incorrect prompt submission
- Technical failures
- Missing response data
- Failure of blinding
- Addition of evidence during claim verification
- Any change to the scoring rules

A protocol deviation will not automatically invalidate the study. Its possible effect on the results will be explained.

33. Limitations

The study is limited to:

- Two AI systems
- One medical topic
- Three prompt conditions
- Three repetitions per condition
- AI versions available during a short testing period
- One primary investigator performing the evidence assessment and scoring

Some subjectivity may remain when classifying partially correct claims or interpreting guideline differences. Blinding may also be imperfect because writing style or citation format could reveal the AI source.

The conclusions will therefore be limited to the systems, prompts, evidence, and testing period examined in this study.

34. Interpretation Rule

The study will not conclude that one AI is universally superior or that all AI-generated medical advice is accurate or inaccurate.

Conclusions will use language such as:

> Under the conditions of this study, prompt framing was—or was not—associated with observable differences in the accuracy, completeness, consistency, or evidentiary support of AI-generated recommendations concerning low-dose aspirin for preeclampsia prevention.